FRONTIERS IN GENETICS

REVIEW ARTICLE
published: 16 December 2013
doi: 10.3389/fgene.2013.00286

# Small proteins: untapped area of potential biological importance

***Mingming Su*[1,2], *Yunchao Ling*[1,2], *Jun Yu*[1], *Jiayan Wu*[1]* *and Jingfa Xiao*[1]***

[1] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China*
[2] *Graduate University of Chinese Academy of Sciences, Beijing, China*

***Edited by:***
*Fengfeng Zhou, Shenzhen Institutes of Advanced Technology, China*

***Reviewed by:***
*Helder I. Nakaya, Emory University, USA*
*Xiyin Wang, Hebei United University, China*

****Correspondence:***
*Jiayan Wu and Jingfa Xiao, Beijing Institute of Genomics, Chinese Academy of Sciences, No.1-7 West Road Bei Chen, Chaoyang District, Beijing 100101, PR China*
*e-mail: wujy@big.ac.cn; xiaojingfa@big.ac.cn*

Polypeptides containing ≤100 amino acid residues (AAs) are generally considered to be small proteins (SPs). Many studies have shown that some SPs are involved in important biological processes, including cell signaling, metabolism, and growth. SP generally has a simple domain and has an advantage to be used as model system to overcome folding speed limits in protein folding simulation and drug design. But SPs were once thought to be trivial molecules in biological processes compared to large proteins. Because of the constraints of experimental methods and bioinformatics analysis, many genome projects have used a length threshold of 100 amino acid residues to minimize erroneous predictions and SPs are relatively under-represented in earlier studies. The general protein discovery methods have potential problems to predict and validate SPs, and very few effective tools and algorithms were developed specially for SPs identification. In this review, we mainly consider the diverse strategies applied to SPs prediction and discuss the challenge for differentiate SP coding genes from artifacts. We also summarize current large-scale discovery of SPs in species at the genome level. In addition, we present an overview of SPs with regard to biological significance, structural application, and evolution characterization in an effort to gain insight into the significance of SPs.

**Keywords: small proteins, small ORFs, protein identification, protein annotation coherence, evolution characterization**

## INTRODUCTION

Proteins generally contain from 50 to 1000 amino acid residues (AAs) per polypeptide chain. In most studies, polypeptides containing ≤100 AAs are considered to be small proteins (SPs) but there is no strict definition of an SP. Some studies have used wider thresholds for SPs of ≤200 AAs (Yang et al., 2011) and some have used narrow thresholds for SPs of ≤85 AAs (Zuber, 2001; Schmidt and Davies, 2007). To date, the smallest protein described is the TAL protein (11 AAs), which influences development of the *Drosophila melanogaster* (Galindo et al., 2007). Because of the short length, SPs generally consist of a simple domain and represent simple, useful model systems for simulation of protein folding (Imperiali and Ottesen, 1999; Polticelli et al., 2001) and for drug design (Martin and Vita, 2000). But many of the earlier studies assumed that the length of a protein sequence is associated with its specific functions and that SPs probably have few notable functions compared to large proteins. According to a statistical survey of SPs, the majority of SPs in a certain species are hypothetical proteins or proteins with unknown functions (Wang et al., 2008), and it is less likely to find shorter proteins with confirmatory homology in other organisms (Lipman et al., 2002; Wang et al., 2008; Zhao et al., 2012). Large proteins have the priority to be annotated (Galperin and Koonin, 2004) and studied while shorter proteins to be relatively unimportant (Hirsh and Fraser, 2001; Jordan et al., 2002). However, the identification of increasing numbers of important SPs has gradually attracted the attention of scientists and many studies have demonstrated that SPs are widespread and have important functionality in all three domains of life (Camby et al., 2006; Galindo et al., 2007; Gleason et al., 2008; Muller et al., 2008; Notaguchi et al., 2008; Oelkers et al., 2008; Jung et al., 2009). In fact, due to binding studies of peptides of various sizes, the minimal size of a functional epitope is ∼8AAs, with an average size of 15–20 AAs. SPs with less than 100 AA are sufficient to contain at least a single domain that exhibits a relevant function or to assist a biological process (Wang et al., 2008). Furthermore, there appears to be a significant evolutionary trend favoring shorter rather than longer proteins for specialized functions (Lipman et al., 2002). This field is receiving increasing interest focused on the significance of SPs. Thus, the bottleneck for the research on SPs might not be the "trivial" functional SPs themselves but the techniques of discovery and analysis of SPs.

## SMALL PROTEIN-CODING GENES OVERLOOKED IN GENOME ANNOTATION

In pace with the rising sequence data in NCBI database, the biggest challenge for whole genome annotation and analysis is becoming to differentiate meaningful gene-coding ORFs from inutile ORFs. Random sequence simulation suggests that, except for long repetitive sequences, ORFs ≥200 AAs are unlikely to occur by chance, whereas a large number of sORFs could include numerous artificial genes (Fickett, 1995; Das et al., 1997). SP-coding genes could easily escape detection in a genome-wide prediction because they are "buried" in an enormous pile of

sORFs (Basrai et al., 1997). Dujon et al. defined a key criterion to annotate an ORF; this criterion takes proteins with ≥100 contiguous codons (including the first ATG) as functional genes and ORFs that are shorter than 100 codons as questionable genes (Dujon et al., 1994). With the application of this criterion, ORFs were identified automatically in the yeast *Saccharomyces cerevisiae* chromosome XI (Dujon et al., 1994). Goffeau also applied this criterion and defined 5885 potential protein-encoding genes from the 12,068 Mb DNA sequence of the *S. cerevisiae* genome, exclusive of SPs (Goffeau et al., 1996). Since then, most algorithms of genome annotation or protein prediction have used a cutoff of ≤100 AAs to reduce the likelihood of false-positive genes. In 2006, Kastenmayer et al. used gene expression-based analyses and homology searching and brought 299 un-annotated sORFs in *S. cerevisiae*, 247 of which have been verified experimentally (Kastenmayer et al., 2006). Thus, the limitations of discovery techniques could have contributed to the assumption that the functions of SPs are less worthy of study. It is suggested that the number of SP-coding genes is substantially greater than those discovered to date, which becomes a challenging problem for biologists trying to predict and validate SPs throughout the genome.

## FUNCTIONAL SIGNIFICANCE OF SPs

As a result of the constraints of discovery techniques, few SPs are identified and the majority of SPs are annotated as hypothetical proteins or proteins with unknown functions. But SPs with known functions are involved in various important functional classes, including information storage and processing, cellular processes and signaling, and metabolism. There is growing evidence that many sORFs in single-celled microorganisms surprisingly encode small bioactive peptides. Some of the well-known SPs include chaperonin, Hsp10, translation initiation factor IF-1, ribosomal proteins and others (Wang et al., 2008). In Bacteria, two SPs, PtrA and PtrB are actively participate in the suppression of the type III secretion system under the stress of DNA damage in *Pseudomonas aeruginosa* (Ha et al., 2004; Wu and Jin, 2005); a group of small acid-soluble spore proteins (SASP) are the crucial factors that protect spore DNA from damaging in dormant spores of *Bacillus*, *Clostridium,* and related species (Setlow, 2007). In yeast, it has been reported that SPs include mating pheromones, proteins involved in energy metabolism, proteolipids, chaperonins, stress proteins, transporters, transcriptional regulators, nucleases, ribosomal proteins, thioredoxins, and metal ion chelators (Basrai et al., 1997). In fact, regulatory and metabolic proteins are more common than constitutive or structural proteins. On the basis of the Clusters of Orthologous Groups (COG) database (Tatusov et al., 2000), we summarized the function types of SPs in Archaea, Bacteria, and Fungi and found that SPs cover nearly all subclasses of functional classes in the COG database, except for constitutive or structural classes; i.e., RNA processing and modification, nuclear structure, and extracellular structures (**Table 1**).

Among multicellular organisms, certain important signaling molecules, hormones, antibacterial defensins, animal toxins, and protease inhibitors belong to the SP family. In plants, some SPs are known to be involved in cell-to-cell communications and regulatory processes. It was demonstrated recently that a membrane-associated thioredoxin (140 AAs) (Meng et al., 2010) is related to intercellular communication, the Cg-1 protein (<33 AAs) (Gleason et al., 2008) controlling the tomato/nematode interaction, the lipid-binding protein AZI1(161 AAs) (Jung et al., 2009) involved in priming plant defenses, and the FLOWERING LOCUS T (FT) protein (175 AAs) (Notaguchi et al., 2008) acting as a long-range signal regulating flowering. In *Arabidopsis*, the CLE family proteins (75–140 AAs) (Fletcher et al., 1999; Trotochaud et al., 2000; Muller et al., 2008) participate in meristem development. CAPRICE (CPC; 94 AAs) is a transcription factor involved in intercellular signal transduction associated with root epidermal cell differentiation (Kurata et al., 2005). In animals, there is a rich diversity of short peptides involved in intercellular transportation and development (Basrai et al., 1997). A eukaryotic TAL protein (11 AA) was reported to influence Drosophila development (Galindo et al., 2007). A long non-coding RNA called polished rice (pri) was found to encode small peptides (11–32 AA) that control proteolytic cleavage of a transcription factor control Shavenbaby (Svb) during *Drosophila* embryogenesis (Kondo et al., 2010). In humans, galectin-1 (135 AAs), for example, plays major roles in neuronal cell differentiation and the establishment and maintenance of T-cell tolerance and homeostasis *in vivo* (Luo et al., 1996). In fact, it is clear that almost all subclasses of functional classes in KOG database (eukaryotic representatives of the COG database) (Tatusov et al., 2003) are covered in *A. thaliana* (*Ath*), *Caenorhabditis elegans* (*Cel*), *Drosophila melanogaster* (*Dme*), *and Homo sapiens* (*Hsa*), except for the nuclear structure class (**Table 1**; see Table S1 for details).

We further studied domains of SPs in the Pfam-A database (Pfam-A) and the NCBI genpept database (NCBI genpept database). The NCBI genpept database contains 14,324,397 proteins, including 1,796,324 (12.54%) SPs. Only 310,909 (17.31%) SPs, about 2.17% of total proteins are annotated, and among the annotated domain SPs, most of them (85.26%) have only one domain (**Figure 1**). SPs usually contain single domain. Domain-known SPs cover 3274 domain items (85.39% of the total 3834 domain items in all Pfam-A families against the NCBI genpept database) (Table S2). Some similar domains are grouped together into clans; we clustered the 3274 domains on the basis of Pfam-C, but not every domain has a corresponding clan. Specifically, 1687 domains belong to clans and 1587 domains were not found in clans (Table S3). Domain analysis revealed that large numbers of SPs are not identified but, as for SPs with known domain, they usually have a simple structure and cover almost all domain classes.

## PREDICTION AND VALIDATION FOR SPs

Although recent advances in computational and experimental approaches make it possible to identify ORFs efficiently at the genome-wide level, there are potential problems for the prediction and validation for SPs.

### HOMOLOGY-BASED SEARCHING

The general technique of homology-based gene prediction is the most reliable tool for discovering evolutionarily conserved genes.

**Table 1 | Function characterization of small proteins in COG/KOG database.**

| COG/KOG function classes | *Archaea* | *Bacteria* | *Fungi* | *Cel* | *Ath* | *Dme* | *Hsa* |
|---|---|---|---|---|---|---|---|
| [J] Translation, ribosomal structure, and biogenesis | + | + | + | + | + | + | + |
| [A] RNA processing and modification | − | − | − | + | + | + | + |
| [K] Transcription | + | + | + | + | + | + | + |
| [L] Replication, recombination, and repair | + | + | − | + | + | + | + |
| [B] Chromatin structure and dynamics | + | + | + | + | + | + | + |
| [D] Cell cycle control, cell division, chromosome partitioning | + | + | + | + | + | + | + |
| [Y] Nuclear structure | − | − | − | − | − | − | − |
| [V] Defense mechanisms | + | + | − | + | − | + | + |
| [T] Signal transduction mechanisms | + | + | + | + | + | + | + |
| [M] Cell wall/membrane/envelope biogenesis | + | + | − | − | + | + | + |
| [N] Cell motility | + | + | − | + | − | + | + |
| [Z] Cytoskeleton | − | + | + | + | + | + | + |
| [W] Extracellular structures | − | − | − | + | + | + | + |
| [U] Intracellular trafficking, secretion, and vesicular transport | + | + | + | + | + | + | + |
| [O] Posttranslational modification, protein turnover, chaperones | + | + | + | + | + | + | + |
| [C] Energy production and conversion | + | + | + | + | + | + | + |
| [G] Carbohydrate transport and metabolism | + | + | + | + | + | − | + |
| [E] Amino acid transport and metabolism | + | + | − | + | + | + | + |
| [F] Nucleotide transport and metabolism | + | + | − | − | + | − | + |
| [H] Coenzyme transport and metabolism | + | + | + | + | + | − | + |
| [I] Lipid transport and metabolism | + | + | + | + | + | + | + |
| [P] Inorganic ion transport and metabolism | + | + | + | + | + | + | + |
| [Q] Secondary metabolites biosynthesis, transport, and catabolism | + | + | − | − | + | + | + |
| [R] General function prediction only | + | + | + | + | + | + | + |
| [S] Function unknown | + | + | + | + | + | + | + |

*"+," found; "−," not found. It describes function types of SPs in Archaea, Bacteria, and Fungi in COG database and those of SPs in eukaryotic species in KOG database. In Archaea, bacteria, and fungi, constitutive or structural classes are not covered, that is, RNA processing and modification, nuclear structure, extracellular structures. In Arabidopsis thaliana (Ath), Caenorhabditis elegans (Cel), Drosophila melanogaster (Dme), and Homo sapiens (Hsa), the nuclear structure class is not covered in all these organisms.*

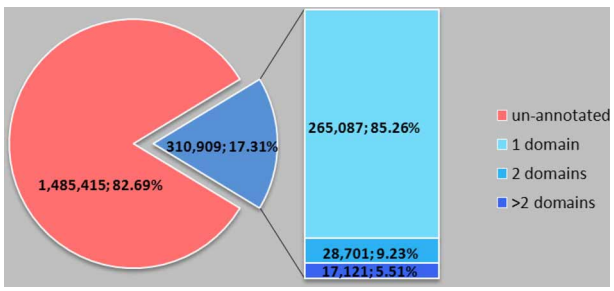


**FIGURE 1 | Domain number distribution of small proteins in NCBI genpept.** SP usually contains a single domain. The NCBI genpept database contains 14,324,397 proteins, including 1,796,324 (12.54%) SPs. Only 310,909 (17.31%) SPs, about 2.17% of total proteins, are annotated, and among the annotated domain SPs, most of them (85.26%) have only one domain.

The key issue is to identify sequence similarity. Each predicted gene from the stringent genome sequence can be annotated if the gene aligns significantly with a known protein sequence from the same organism or other organisms. It has been used to discover small, un-annotated, protein-coding, and non-protein-coding genes in chromosomal regions previously considered to be intergenic region between the genome of *S. cerevisiae* and those of other hemiascomycetous yeasts (Blandin et al., 2000) or other Saccharomyces genomes (Cliften et al., 2001). However, as for SPs, the sequence similarity-based homology assessment method is limited by the large size of the protein database and the short length of the sequence. The expectation (*e*-value) for finding a random sequence match in a database often takes into account the length of the sequence and, for a short query sequence, the probability of a random match is quite high. Thus, alignment-based methods exclude a lot of small potential genes, which are often classified as ORFs that occur by chance (Skovgaard et al., 2001). Moreover, it is also less likely to find SPs that do not have confirmatory homology in other organisms. By BLAST comparisons, Wang F et al. took several representative phyla to investigate conservation among SPs and found the species-specific SPs are the majority in all of the phyla (Wang et al., 2008). Thus, homology searching is not competent for discovery of species-specific SPs. And SPs like ubiquitin (76 AA, not including the pre-protein peptide), which is highly conserved from fungi to mammals but only sharing a similar structure with ubiquitin proteins in prokaryotic cells, are also excluded by alignment-based approach (Bienkowska et al., 2003).

### PURELY STATISTICAL ALGORITHMS

Purely statistical algorithms make gene prediction from multiple features of gene-coding sequences. Methods on purely statistical grounds, using either probabilistic or pattern-based schemes to score candidate genes, display high sensitivity for discovering genes without a match. Most *ab initio* gene prediction programs distinguish coding (CDS) and non-coding sequences (NCDS) with their differences in nucleotide composition, intron splice sites, promoters, translational start/stop sites, and polyadenylation signals. These signals are generally integrated for evaluating the coding likelihood of a sequence. However, SPs prediction tools could not take all the characteristics into consideration. The integration of multiple criteria decreases the chance that false exons are predicted as true (low false-positive rate) but likely increases the chance that true exons are not predicted (high false-negative rate). The issue of false-negative prediction is particularly serious for smaller CDSs ($\leq$300 nucleotides) due to the difficulty in distinguishing the relatively few biologically meaningful sequences from the very large pool of small ORFs (sORFs) (Basrai et al., 1997). Noting the relatively high false-negative rate of current gene finding algorithms and the difficulty to identify SP genes, recent studies focus only one or two signals to predict sORFs and they also take considerations from functional constraints in the follow-up analysis. Hanada et al. developed the program package sORF finder for identifying sORFs according to the nucleotide composition bias among coding sequences and the potential functional constraint at the AA level through evaluation of KA/KS ratio, because a functional coding sORF is expected to undergo stronger selective constraints on non-synonymous sites than for synonymous ones. Yang XH et al. identified candidate sORFs set by protein domain-scanning that is searching the InterPro database for annotated protein domain/motifs (Yang et al., 2011). The codon adaptation index (CAI), which is based on the similarity of usage of preferred and a limited number of codons for highly expressed genes, has been used to evaluate the coding potential of a putative ORF (Sharp and Li, 1987). Hanada et al. used this simple gene-finding method in a large-scale search for sORFs encoding proteins of 30–100 AA in the intergenic regions of the Arabidopsis genome (Hanada et al., 2007). On the basis of this research, Hanada et al. developed the program package sORF finder for identifying sORFs according to the nucleotide composition bias among coding sequences and the potential functional constraint at the AA level through evaluation of synonymous and non-synonymous substitution rates (Hanada et al., 2010). This measurement becomes less robust for sORFs, because the parameter will fluctuate dramatically as ORF length decreases. Only 2% of un-annotated sORFs predicted by Hanada et al. (2007) were confirmed by the Arabidopsis proteomic data (Castellana et al., 2008). Later, Termier and Kalogeropoulos examined the probability of functionality of sORFs and described computational techniques based on a combination of codon usage, AA composition, and dipeptide frequencies in the encoded protein to distinguish coding and non-coding sequences (Termier and Kalogeropoulos, 1996). In addition, Xiaohan Yang et al. reported an integrative sORF discovery strategy based on transcriptomics, proteomics, and computational biology, which was validated by both bioinformatics (e.g., protein domain-scanning) and experimental approaches (e.g., protein mass spectrometry) (Yang et al., 2011). Although multiple criteria could minimize the risk of considering a fortuitous ORF to be a meaningful protein-coding gene, it is also hardly to achieve efficient designation of small coding sequences from the very large pool of sORFs because of the growing error of short sequence annotation.

### EVIDENCE-BASED STRATEGIES

Many experimental strategies have been used as gene prediction or validation tools and these methods have the ability to predict novel genes that could not be identified *in silico*. Most of them are based on gene expression data, such as RNA-Seq, EST (Expressed Sequence Tags), DNA microarray, and SAGE (serial analysis of gene expression). Although expression cannot (at all) be used to validate the translation of a SP, it is still the effective approach to address a SP candidate. Recent studies show a squared Pearson correlation coefficient of $\sim$0.40, which implies that $\sim$40% of the variation in protein concentration can be explained by knowing mRNA abundances (Vogel and Marcotte, 2012). Yamada et al. have used *Arabidopsis thaliana* full-length cDNA data and EST data from *A. thaliana*, *Brassica*, rice, and wheat to pinpoint transcribed, un-annotated genomic regions to identify novel transcribed sequences in *A. thaliana* (Yamada et al., 2003). Each plausible gene can be identified if it matches with EST or cDNA sequence. But some coding sORFs may be either expressed under specific conditions not covered or tend to have significantly lower expression levels than long high expressional genes leading to few evidence of sORFs to be found in transcriptome experiments. Another approach that has been developed is a microarray-based method, which is often used as a gene validation tool. The core principle behind microarrays is hybridization between two DNA strands. A single "chip" or array contains probes to determine transcript levels for every known gene in the genome of one or more organisms simultaneously. Shoemaker et al. used microarrays to refine and validate computational gene predictions for the human genome and define full-length transcripts on the basis of co-regulated expression of their exons (Shoemaker et al., 2001). It can provide more accurate results and represents a powerful tool for identifying transcripts. Nevertheless, this method requires explicitly designed chips and some small transcripts might not be systematically defined to allow the creation of the required chips. The serial analysis of gene expression (SAGE) technique can provide quantitative gene expression data without the prerequisite of a hybridization probe for each transcript. The general goal of SAGE technique is similar to DNA microarray and the difference between SAGE and microarrays is that SAGE sampling is based on lists of short sequence tags, not on hybridization of mRNA output to probes. The tag-based gene expression profiling can measure the expression levels of known or unknown sequences. It has been designed to catalog transcripts including a small number of unpredicted sORFs on a genome-wide level in yeast genome studies (Basrai et al., 1997; Velculescu et al., 1997; Basrai et al., 1999). Although it has the advantage of greater sensitivity to low levels of expression, the number of sORFs identified will be limited by the number of tags analyzed, the physiological state from which they are isolated, and the restriction enzyme used to define tags (Basrai et al., 1997).

In addition to traditional transcriptional methods mentioned above, next-generation sequencing refreshes the methodology of transcriptomics that is, it directly sequences transcriptomes. By using deep sequencing technologies to sequence cDNA, RNA-Seq has been developed to transcriptome profiling quantitatively (Wang et al., 2009). The expression levels determined by RNA-Seq, which does not suffer from problems with background noise, are more accurate than traditional cross-hybridization methods. If the sequencing depth is sufficient, RNA-Seq would discover novel transcripts especially for SPs, some of which are hardly to be detected effectively in traditional expression-based methods. Despite of the individual advantages and limitations, all of the expression-based methods have several potential problems to identify SPs because sensitivity is contingent on the extent of the expression datasets, which might exclude genes with little evidence or expressed in uncovered specific conditions. For example, only low-level expressors of the SP of DAP-5 could be selected by the transfections with the original episomal-based vector or with the bicistronic vector, because overexpression of the DAP-5 was lethal to HeLa cells (Levy-Strumpf et al., 1997). Another SP is negative p53, whose expression inhibits DEK RNA interference-induced p53 transcriptional induction, as well as cell death, thus directly implicating p53 activation in the observed apoptotic phenotype (Wise-Draper et al., 2006). Therefore, those low-level expressed SPs are difficult to be verified by expression-based methods. These problems would not be encountered by analysis of a collection of transposon insertions, which can identify genes expressed at different times in the life cycle and determine the subcellular locations of the encoded gene products as well as the phenotype of the disrupted strains. No cDNA is required and the majority of new genes are either short or overlap a previously un-notated gene on the opposite strand. Smith et al. described a genetic footprinting method based on the endogenous yeast transposon Ty1 (Smith et al., 1996). This method could be useful for identifying sORFs if primers against interfeature regions (regions lying between known ORFs, tRNA genes, or other sequence "features"). In gene-finding studies in yeast (Ross-Macdonald et al., 1999; Kumar et al., 2002), candidate genes are identified by means of large-scale shuttle mutagenesis (Seifert et al., 1986) with a modified transposon as a simple gene trap. However, genes encoding proteins below a 100 AA cutoff might be under-represented in a mutational search because of the small target size for mutagenesis and the number of insertions analyzed. In addition to the expression-based method and mutagenesis, mass spectrometry has allowed for large-scale surveys of the proteome. Yang XH et al. identified highest-confidence candidate sORFs set by proteomics data using protein mass spectrometry. Proteomics has now advanced sufficiently to allow for the systematic quantification of proteins. But it also excludes large amounts of protein-coding sORFs because of the fast and dynamic nature of biological process.

## INTEGRATED STRATEGIES

As for the prediction of short protein-coding genes, the challenge is that short non-coding ORFs are difficult to distinguish from real genes; the shorter the protein, the greater the probability of error rate of detection. No single technique is comprehensive. In order to predict short genes completely and correctly, most studies combine both genome-wide searching algorithms *in silico* and expression analysis. Kumar et al. integrated methods of gene-trapping, microarray-based expression analysis, and genome-wide homology searching for finding overlooked sORFs and antisense sORFs in yeast (Ross-Macdonald et al., 1999; Kumar et al., 2002). The 137 genes discovered using this approach, including 104 SPs-coding genes, constitute 2% of the yeast genome and represent a wealth of overlooked biology. Yang et al. reported an integrative sORFs discovery strategy based on experimental data (transcriptome), coding potential prediction, evolutionary conservation, and gene family clustering (Yang et al., 2011). The sORF candidates predicted in this study display a relatively high rate of proteomics support and protein mass spectrometry support. We provided an overview of the integrated strategies for SPs prediction (**Figure 2**). In the computational stage, *in silico* programs could discover rarely expressed sORFs or tightly regulated sORFs, which are hardly detected in experimental methods. Homology searching methods are valuable for conserved SPs discovery but are not available for novel SPs candidates and some SPs share similar structure. Pure statistically algorithms combine multiple parameters and generate feature or pattern of SPs, which could be conducted from training set of SPs-coding genes in relative organisms. It is efficient for the designation of non-conserved small coding genes excluded by alignment-based methods. But computational methods include many false positives, some of which could be validation in the experimental stage. Expression-based approaches directly assess

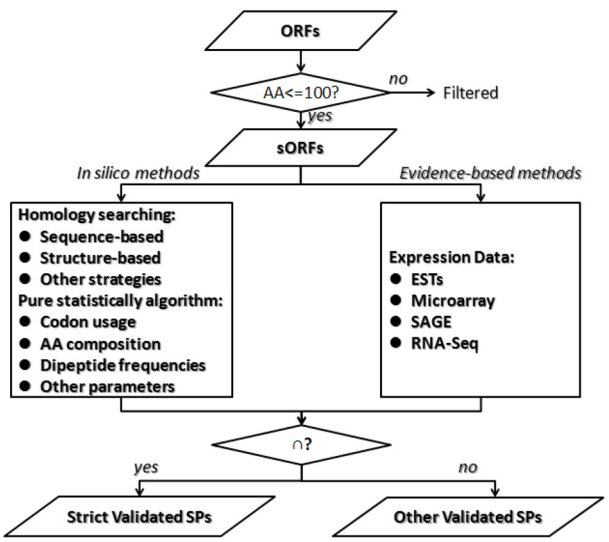


**FIGURE 2 | An overview of integrated strategies for small proteins prediction.** It is challenge to differentiate meaningful gene-coding sORFs from inutile sORFs because the shorter the protein sequence, the greater the probability of error rate of detection. First we suggest splitting the annotation of SPs from other proteins. Second, it is better to combine both *in silico* algorithms and evidence-based analysis. Then merge the two parts of results and get two sets of SPs as follows. The strict validated SPs are those validated by both methods, while other validated SPs are those only validated by either *in silico* algorithms or evidence-based analysis.

the gene expression level, which supplement to validate the meaningfulness of predicted sORFs. Experimental methods could also predict some new sORFs missed in computational stage. This integrated strategy could improve the sensitivity and specificity of annotating SPs-coding genes in stringent genomes, but the choice of each method applied and each parameter set is contingent.

## GENOME-WIDE PREDICTION IN SPECIES LEVEL

Because of the advances in experimental and computational approaches, it has emerged that most studies (**Table 2**) are focused on large-scale discovery of SPs in species rather than a small number of SP families in specific organisms. In prokaryotes, Peter performed systematic function analysis of potential proteins, including 345 small polypeptide ORFs (of 85 codons or less) in *Bacillus subtilis,* which is known to produce an abundance of small polypeptides (Zuber, 2001). In single-cell eukaryotes, e.g., *S. cerevisiae*, Marco used genome-wide comparative analysis and identified 117 novel small genes, 84 of which are transcribed (Kessler et al., 2003). Kastenmayer et al. used gene expression-based analyses and homology searching and brought the total number of un-annotated sORFs in *S. cerevisiae* to 299, 247 of which have been verified experimentally (Kastenmayer et al., 2006). Earlier studies in plants showed that relatively little is known about sORF genes except for a number of small secreted proteins in *A. thaliana* (Cock and McCormick, 2001; Butenko et al., 2003). Recent studies have revealed a large number of novel coding sORFs. Lease and Walker predicted 33,809 unannotated *Arabidopsis* ORFs encoding SPs of 25–250 AA in length (Lease and Walker, 2006); Hanada identified 3241 coding sORFs with either evidence of transcription or purifying selection which likely to be novel coding gene (Hanada et al., 2007). In animals, Emmanuel and Vini identified nearly 600,000 sORFs in the putatively non-coding euchromatic DNA of *Drosophila melanogaster* (Ladoukakis et al., 2011); Frith et al. reported that ∼10% of proteins in *Mus musculus* are <100 AAs, although the majority of these are variants of proteins that are >100 AAs (Frith et al., 2006). Despite the inherent difficulties of identifying sORFs, these publications of large-scale discovery efforts may reveal additional sORFs with more valuable data and more advanced sORFs discovery methods.

## STRUCTURAL APPLICATION FOR SPs

Except for functional importance, many studies have demonstrated that SPs containing <40 AAs with a compact, folded structure provide simple model systems for studying protein folding and stability as well as serving as scaffolds for the rational design of new functional motifs (Cunningham and Wells, 1997; DeGrado et al., 1999; Imperiali and Ottesen, 1999), which benefits both computational simulation and pharmaceutical studies.

In the simulation of protein folding, SPs are often used as model systems to overcome folding speed limits and to provide insight into the complex architecture of proteins. Generally, the polypeptide chains that are made up of thousands of atoms and hence consist of millions of possible interatomic interactions. It might be supposed that the resulting complexity would make the accurate prediction of protein structure and protein-folding mechanisms nearly impossible (Baker, 2000). However, SPs and domains can be folded quickly and correctly as the number of factors that influence folded state stability is reduced. As a result, many studies have used small motifs for structural simulation. Struthers et al. showed a metal-independent folded structure (ββα) reproduced in a 23 AA peptide through an iterative process (Struthers et al., 1996). Jennifer et al. designed a discretely folded SP motif based on the toxin hand (TH) motifs (Ottesen and Imperiali, 2001). Neidigh et al. have reported the smallest stable structural Trp-cage motif, a 20 AA peptide that adopts a well-defined globular shape, which provides a new tool for elucidating protein conformational preferences (Gellman and Woolfson, 2002; Neidigh et al., 2002; Qiu et al., 2002). The SP motif rapidly and accurately provides an excellent model for secondary structure simulation and provides the foundation for understanding the structures of large proteins.

The engineering of novel functional SPs has the potential to become a fundamental step toward the conversion of a protein functional epitope or a flexible peptide lead into a classical pharmaceutical. Such SPs represent a potential intermediate step in the development of drugs targeted to a protein–protein interface (Cunningham and Wells, 1997). The design of bioactive small molecules for interaction at large protein–protein interfaces remains a challenge and many studies are focused on minimizing proteins into significantly smaller polypeptides via both rational design processes and selection from vast combinational libraries

**Table 2 | Summary of large-scale sORF studies in different organisms.**

| | Organism | Genome size (Mbp) | Protein-coding genes | sORFs[a] | Verified[b] | %[c] | Source |
|---|---|---|---|---|---|---|---|
| Prokaryotes | *Bacillus subtilis* | 4 | 4100 | 345 | 180 | 4 | *Peptides*, 2001. 22(10) |
| Eukaryotes | *Saccharomyces cerevisiae* | 12 | 5865 | 299 | 247 | 4 | *Genome Res*, 2003. 13(2); *Genome Res*, 2006. 16(3) |
| | *Arabidopsis thaliana* | 120 | 29,157 | 7159 | 3241 | 11 | *Plant Physiol*, 2006. 142(3); *Genome Res*, 2007. 17(5) |
| | *Drosophila melanogaster* | 180 | 13,907 | 4561 | 401 | 3 | *Genome Biol*, 2011. 12(11) |
| | *Mus musculus* | 2500 | 31,035 | 1240 | 1167 | 4 | *PLoS Genet*, 2006. 2(4) |

*It describes studies focused on large-scale discovery of SPs in species and their results. [a]Numbers of coding or annotated sORFs (<100 AA); [b]Numbers of sORFs with experimental evidence or known function; [c]The fraction of verified sORFs relative to previously annotated protein coding genes.*

(Martin and Vita, 2000). To date, scientists have designed a few SPs whose stability or instability has enhanced our understanding of those rules. Both of the natural (e.g., α/β scorpion toxin fold, protease inhibitors, leucine zipper, and zinc finger) and artificial SPs (TASP) have been used as structural scaffolds in the engineering of novel binding activity (Martin and Vita, 2000). Some of them can be used directly in therapy or exhibit a high potential to serve as drugs. In all cases, they represent precious structural intermediates that are useful as identification frameworks for peptidomimetic design or lead directly to new small organic structures, representing novel drug candidates.

## EVOLUTION CHARACTERIZATION OF SPs

Proteins evolve under a variety of constraints, for example, as specific functions, base or AA compositions (Knight et al., 2001) and sequence length (Lipman et al., 2002). Studies of the evolutionary characterization of SPs draw attention to the question of how evolutionary trends affect variation of protein length. Two obvious observations from the evolutionary characterization of SPs are as follows. First, SPs are likely to change whereas long proteins are likely to be conserved. Studies (Guigo et al., 2003; Wei et al., 2005; Windsor and Mitchell-Olds, 2006) indicate that computational gene prediction methods are not generally capable of identifying SPs, which display elevated Ka/Ks ratios in interspecific comparisons, suggesting that SPs are generally rapidly evolving sequences. Furthermore, Lipman et al. studied the relationship between length and conservation (Lipman et al., 2002) and Zhao et al. analyzed SPs across eight Eukaryotes (Zhao et al., 2012). It is found that SPs tend to be non-conventional proteins and appear to have lineage-specific or tissue-specific function. There appears to be a significant evolutionary trend favoring shorter rather than longer proteins, possibly because of the need to minimize the cost of protein translation and the cost of the relationships that are required to fold longer, particularly multi-domain, proteins (Hartl and Hayer-Hartl, 2002). Perhaps too many changes in longer proteins would increase the risk of undesirable side-effects; i.e., deleterious interactions with other cellular components. The evolutionarily stable core of archaeal genomes includes the great majority of genes coding for conserved proteins involved in genome replication and expression, but only a relatively small subset of metabolic functions (Makarova et al., 1999). By contrast, the majority of SPs involve metabolic processes, transcriptional regulation or cell communication rather than essential roles in organisms. It is possible that vital functional proteins are more conserved than regulatory proteins in order to decrease side-effects, whereas poorly conserved proteins appear to tend toward minimal domain size and retain lineage-specific functions. Second, SPs are ancient and the origin of the protein universe is highly likely to have arisen from SPs with simple hydrogen-bonded, secondary structural elements instead of the details of side-chains. There is a tendency toward greater protein length along with increasingly complex genomes. Many prokaryotes generally have shorter proteins, on average, than eukaryotes (Makarova et al., 1999). Among the eukaryotes, proteins of the microsporidium *Encephalitozoon cuniculi*, which has an extremely compact genome, are smaller than the corresponding proteins in organisms with larger genomes (Katinka et al., 2001). There is evidence that a very small set of secondary structural elements, compacted from non-homologous representative proteins in the Protein Data Bank (PDB) of 41–150 residues, is complete for single-domain protein structures (Zhang et al., 2006). Similarly, we found that a very small set of SPs in the NCBI genpept cover a large number of domains in Pfam-A families (Table S2). These results support that SPs contain the majority, if not all, of the core secondary structural element, which can be used as the starting template. As SPs evolve, some could be folded into compact multi-domain proteins, whereas others could prefer to remain as small as originally created; at the same time, some new proteins are created along with species differentiation.

Above all, these two observations suggest that SPs might have important roles in evolutionary trends and give an possible answer to why nature needs SPs, but some intriguing questions that remain unanswered are focused on what a unique evolutionary pattern of SPs is and how an SP could reveal additional surprises.

## CONCLUSIONS AND PERSPECTIVES

SPs generally consist of a simple domain and tend to be treated as trivial molecules in biological processes. Large proteins have become priority targets to be analyzed whereas study of SPs is an almost untapped virgin territory in biological research. Despite an increasing number of SPs to be identified and involved in various biological functions, the vast majority of SPs are annotated as hypothetical proteins or proteins with function unknown. This is partly due to the limitations and challenges in most current gene discovery techniques, which are not generally appropriate for SPs identification. SPs have largely escaped detection and are hard to be differentiated from large amounts of artifacts. The integrated strategies combing *in silico* algorithms and evidence-based analysis could be more capable to discover potential SPs to some extent, although these strategies require improvements and it is also required more specific algorithms and techniques in this aspect produced in the future. Recent detection success suggests it is possible for large-scale identification and systematic analysis of SPs or sORFs at the genome level instead of only a limited number of SP families. The more exciting thing is scientists are gradually paying more attention toward solving exciting questions, such as why does nature need SPs if there are functional characterizations or unique evolutionary patterns for small peptides? The same question might arise for micro RNAs or small RNAs. Besides, smaller motifs have length advantages in iterative modeling, synthesis and structural characterization, prompting interest in discovering efficient testing procedures for pharmaceutical design strategies or principles.

## ACKNOWLEDGMENTS

The authors are grateful to the International Science Editing for language editing and making this a better review article and Dr. Zhong Jin of Supercomputing Center, Computer Network Information Center, Chinese Academy of Sciences for critical reading and constructive suggestions.

## FUNDING

This work was supported by the National Basic Research Program (973 Program) [2010CB126604]; the Special Foundation Work Program [2009FY120100]; the National Programs for High Technology Research and Development (863 Program) [2012AA020409]; the Ministry of Science and Technology of the People's Republic of China; and the National Science Foundation of China [31071163].

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.00286/abstract

## REFERENCES

Baker, D. (2000). A surprising simplicity to protein folding. *Nature* 405, 39–42. doi: 10.1038/35011000

Basrai, M. A., Hieter, P., and Boeke, J. D. (1997). Small open reading frames: beautiful needles in the haystack. *Genome Res.* 7, 768–771. doi: 10.1101/gr.7.8.768

Basrai, M. A., Velculescu, V. E., Kinzler, K. W., and Hieter, P. (1999). NORF5/HUG1 is a component of the MEC1-mediated checkpoint response to DNA damage and replication arrest in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 19, 7041–7049.

Bienkowska, J. R., Hartman, H., and Smith, T. F. (2003). A search method for homologs of small proteins. Ubiquitin-like proteins in prokaryotic cells? *Protein Eng.* 16, 897–904. doi: 10.1093/protein/gzg130

Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited. *FEBS Lett.* 487, 31–36. doi: 10.1016/S0014-5793(00)02275-4

Butenko, M. A., Patterson, S. E., Grini, P. E., Stenvik, G. E., Amundsen, S. S., Mandal, A., et al. (2003). Inflorescence deficient in abscission controls floral organ abscission in Arabidopsis and identifies a novel family of putative ligands in plants. *Plant Cell* 15, 2296–2307. doi: 10.1105/tpc.014365

Camby, I., Le Mercier, M., Lefranc, F., and Kiss, R. (2006). Galectin-1: a small protein with major functions. *Glycobiology* 16, 137R–157R. doi: 10.1093/glycob/cwl025

Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S. P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl. Acad. Sci. U.S.A.* 105, 21034–21038. doi: 10.1073/pnas.0811066106

Cliften, P. F., Hillier, L. W., Fulton, L., Graves, T., Miner, T., Gish, W. R., et al. (2001). Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* 11, 1175–1186. doi: 10.1101/gr.182901

Cock, J. M., and McCormick, S. (2001). A large family of genes that share homology with CLAVATA3. *Plant Physiol.* 126, 939–942. doi: 10.1104/pp.126.3.939

Cunningham, B. C., and Wells, J. A. (1997). Minimized proteins. *Curr. Opin. Struct. Biol.* 7, 457–462. doi: 10.1016/S0959-440X(97)80107-8

Das, S., Yu, L., Gaitatzes, C., Rogers, R., Freeman, J., Bienkowska, J., et al. (1997). Biology's new Rosetta stone. *Nature* 385, 29–30. doi: 10.1038/385029a0

DeGrado, W. F., Summa, C. M., Pavone, V., Nastri, F., and Lombardi, A. (1999). *De novo* design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* 68, 779–819. doi: 10.1146/annurev.biochem.68.1.779

Dujon, B., Alexandraki, D., Andre, B., Ansorge, W., Baladron, V., Ballesta, J. P., et al. (1994). Complete DNA sequence of yeast chromosome XI. *Nature* 369, 371–378. doi: 10.1038/369371a0

Fickett, J. W. (1995). ORFs and genes: how strong a connection? *J. Comput. Biol.* 2, 117–123. doi: 10.1089/cmb.1995.2.117

Fletcher, J. C., Brand, U., Running, M. P., Simon, R., and Meyerowitz, E. M. (1999). Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. *Science* 283, 1911–1914. doi: 10.1126/science.283.5409.1911

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., et al. (2006). The abundance of short proteins in the mammalian proteome. *PLoS Genet.* 2:e52. doi: 10.1371/journal.pgen.0020052

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., and Couso, J. P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* 5:e106. doi: 10.1371/journal.pbio.0050106

Galperin, M. Y., and Koonin, E. V. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 32, 5452–5463. doi: 10.1093/nar/gkh885

Gellman, S. H., and Woolfson, D. N. (2002). Mini-proteins Trp the light fantastic. *Nat. Struct. Biol.* 9, 408–410. doi: 10.1038/nsb0602-408

Gleason, C. A., Liu, Q. L., and Williamson, V. M. (2008). Silencing a candidate nematode effector gene corresponding to the tomato resistance gene Mi-1 leads to acquisition of virulence. *Mol. Plant Microbe Interact.* 21, 576–585. doi: 10.1094/MPMI-21-5-0576

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H. et al. (1996). Life with 6000 genes. *Science* 274, 546, 563–567.

Guigo, R., Dermitzakis, E. T., Agarwal, P., Ponting, C. P., Parra, G., Reymond, A., et al. (2003). Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci. U.S.A.* 100, 1140–1145. doi: 10.1073/pnas.0337561100

Ha, U. H., Kim, J., Badrane, H., Jia, J., Baker, H. V., Wu, D., et al. (2004). An *in vivo* inducible gene of *Pseudomonas aeruginosa* encodes an anti-ExsA to suppress the type III secretion system. *Mol. Microbiol.* 54, 307–320. doi: 10.1111/j.1365-2958.2004.04282.x

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., and Shiu, S. H. (2010). sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* 26, 399–400. doi: 10.1093/bioinformatics/btp688

Hanada, K., Zhang, X., Borevitz, J. O., Li, W. H., and Shiu, S. H. (2007). A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Res.* 17, 632–640. doi: 10.1101/gr.5836207

Hartl, F. U., and Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* 295, 1852–1858. doi: 10.1126/science.1068408

Hirsh, A. E., and Fraser, H. B. (2001). Protein dispensability and rate of evolution. *Nature* 411, 1046–1049. doi: 10.1038/35082561

Imperiali, B., and Ottesen, J. J. (1999). Uniquely folded mini-protein motifs. *J. Pept. Res.* 54, 177–184. doi: 10.1034/j.1399-3011.1999.00121.x

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12, 962–968. doi: 10.1101/gr.87702

Jung, H. W., Tschaplinski, T. J., Wang, L., Glazebrook, J., and Greenberg, J. T. (2009). Priming in systemic plant immunity. *Science* 324, 89–91. doi: 10.1126/science.1170025

Kastenmayer, J. P., Ni, L., Chu, A., Kitchen, L. E., Au, W. C., Yang, H., et al. (2006). Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res.* 16, 365–373. doi: 10.1101/gr.4355406

Katinka, M. D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., et al. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414, 450–453. doi: 10.1038/35106579

Kessler, M. M., Zeng, Q., Hogan, S., Cook, R., Morales, A. J., and Cottarel, G. (2003). Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. *Genome Res.* 13, 264–271. doi: 10.1101/gr.232903

Knight, R. D., Freeland, S. J., and Landweber, L. F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* 2, RESEARCH0010. doi: 10.1186/gb-2001-2-4-research0010

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., et al. (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science* 329, 336–339. doi: 10.1126/science.1188158

Kumar, A., Harrison, P. M., Cheung, K. H., Lan, N., Echols, N., Bertone, P., et al. (2002). An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.* 20, 58–63. doi: 10.1038/nbt0102-58

Kurata, T., Ishida, T., Kawabata-Awai, C., Noguchi, M., Hattori, S., Sano, R., et al. (2005). Cell-to-cell movement of the CAPRICE protein in Arabidopsis root epidermal cell differentiation. *Development* 132, 5387–5398. doi: 10.1242/dev.02139

Ladoukakis, E., Pereira, V., Magny, E., Eyre-Walker, A., and Couso, J. P. (2011). Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol* 12, R118. doi: 10.1186/gb-2011-12-11-r118

Lease, K. A., and Walker, J. C. (2006). The Arabidopsis unannotated secreted peptide database, a resource for plant peptidomics. *Plant Physiol.* 142, 831–838. doi: 10.1104/pp.106.086041

Levy-Strumpf, N., Deiss, L. P., Berissi, H., and Kimchi, A. (1997). DAP-5, a novel homolog of eukaryotic translation initiation factor 4G isolated as a putative modulator of gamma interferon-induced programmed cell death. *Mol. Cell. Biol.* 17, 1615–1625.

Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R., and Tatusova, T. A. (2002). The relationship of protein conservation and sequence length. *BMC Evol. Biol.* 2:20. doi: 10.1186/1471-2148-2-20

Luo, Z., Wang, R., and Lai, L. (1996). RASSE: a new method for structure-based drug design. *J. Chem. Inf. Comput. Sci.* 36, 1187–1194. doi: 10.1021/ci950277w

Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I., et al. (1999). Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* 9, 608–628.

Martin, L., and Vita, C. (2000). Engineering novel bioactive mini-proteins from small size natural and de novo designed scaffolds. *Curr. Protein Pept. Sci.* 1, 403–430. doi: 10.2174/1389203003381306

Meng, L., Wong, J. H., Feldman, L. J., Lemaux, P. G., and Buchanan, B. B. (2010). A membrane-associated thioredoxin required for plant growth moves from cell to cell, suggestive of a role in intercellular communication. *Proc. Natl. Acad. Sci. U.S.A.* 107, 3900–3905. doi: 10.1073/pnas.0913759107

Muller, R., Bleckmann, A., and Simon, R. (2008). The receptor kinase CORYNE of Arabidopsis transmits the stem cell-limiting signal CLAVATA3 independently of CLAVATA1. *Plant Cell* 20, 934–946. doi: 10.1105/tpc.107.057547

NCBI genpept database. Available online at: ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam26.0/ncbi.gz.

Neidigh, J. W., Fesinmeyer, R. M., and Andersen, N. H. (2002). Designing a 20-residue protein. *Nat. Struct. Biol.* 9, 425–430. doi: 10.1038/nsb798

Notaguchi, M., Abe, M., Kimura, T., Daimon, Y., Kobayashi, T., Yamaguchi, A., et al. (2008). Long-distance, graft-transmissible action of Arabidopsis FLOWERING LOCUS T protein to promote flowering. *Plant Cell Physiol.* 49, 1645–1658. doi: 10.1093/pcp/pcn154

Oelkers, K., Goffard, N., Weiller, G. F., Gresshoff, P. M., Mathesius, U., and Frickey, T. (2008). Bioinformatic analysis of the CLE signaling peptide family. *BMC Plant Biol.* 8:1. doi: 10.1186/1471-2229-8-1

Ottesen, J. J., and Imperiali, B. (2001). Design of a discretely folded mini-protein motif with predominantly beta-structure. *Nat. Struct. Biol.* 8, 535–539. doi: 10.1038/88604

Pfam-A. Available online at: ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam26.0/Pfam-A.full.ncbi.gz.

Polticelli, F., Raybaudi-Massilia, G., and Ascenzi, P. (2001). Structural determinants of mini-protein stability. *Biochem. Mol. Biol. Educ.* 29, 16–20. doi: 10.1016/S1470-8175(00)00066-7

Qiu, L., Pabit, S. A., Roitberg, A. E., and Hagen, S. J. (2002). Smaller and faster: the 20-residue Trp-cage protein folds in 4 micros. *J. Am. Chem. Soc.* 124, 12952–12953. doi: 10.1021/ja0279141

Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., et al. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* 402, 413–418. doi: 10.1038/46558

Schmidt, E. E., and Davies, C. J. (2007). The origins of polypeptide domains. *Bioessays* 29, 262–270. doi: 10.1002/bies.20546

Seifert, H. S., Chen, E. Y., So, M., and Heffron, F. (1986). Shuttle mutagenesis: a method of transposon mutagenesis for *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* 83, 735–739. doi: 10.1073/pnas.83.3.735

Setlow, P. (2007). I will survive: DNA protection in bacterial spores. *Trends Microbiol* 15, 172–180. doi: 10.1016/j.tim.2007.02.004

Sharp, P. M., and Li, W. H. (1987). The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295. doi: 10.1093/nar/15.3.1281

Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., et al. (2001). Experimental annotation of the human genome using microarray technology. *Nature* 409, 922–927. doi: 10.1038/35057141

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., and Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* 17, 425–428. doi: 10.1016/S0168-9525(01)02372-1

Smith, V., Chou, K. N., Lashkari, D., Botstein, D., and Brown, P. O. (1996). Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* 274, 2069–2074. doi: 10.1126/science.274.5295.2069

Struthers, M. D., Cheng, R. P., and Imperiali, B. (1996). Design of a monomeric 23-residue polypeptide with defined tertiary structure. *Science* 271, 342–345. doi: 10.1126/science.271.5247.342

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41. doi: 10.1186/1471-2105-4-41

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

Termier, M., and Kalogeropoulos, A. (1996). Discrimination between fortuitous and biologically constrained open reading frames in DNA sequences of *Saccharomyces cerevisiae*. *Yeast* 12, 369–384. doi: 10.1002/(SICI)1097-0061(19960330)12:4&<369::AID-YEA922&>3.0.CO;2-#

Trotochaud, A. E., Jeong, S., and Clark, S. E. (2000). CLAVATA3, a multimeric ligand for the CLAVATA1 receptor-kinase. *Science* 289, 613–617. doi: 10.1126/science.289.5479.613

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr. et al. (1997). Characterization of the yeast transcriptome. *Cell* 88, 243–251. doi: 10.1016/S0092-8674(00)81845-0

Vogel, C., and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232. doi: 10.1038/nrg3185

Wang, F., Xiao, J., Pan, L., Yang, M., Zhang, G., Jin, S., et al. (2008). A systematic survey of mini-proteins in bacteria and archaea. *PLoS ONE* 3:e4027. doi: 10.1371/journal.pone.0004027

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M., et al. (2005). Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res.* 15, 577–582. doi: 10.1101/gr.3329005

Windsor, A. J., and Mitchell-Olds, T. (2006). Comparative genomics as a tool for gene discovery. *Curr. Opin. Biotechnol.* 17, 161–167. doi: 10.1016/j.copbio.2006.01.007

Wise-Draper, T. M., Allen, H. V., Jones, E. E., Habash, K. B., Matsuo, H., and Wells, S. I. (2006). Apoptosis inhibition by the human DEK oncoprotein involves interference with p53 functions. *Mol. Cell. Biol.* 26, 7506–7519. doi: 10.1128/MCB.00430-06

Wu, W., and Jin, S. (2005). PtrB of *Pseudomonas aeruginosa* suppresses the type III secretion system under the stress of DNA damage. *J. Bacteriol.* 187, 6058–6068. doi: 10.1128/JB.187.17.6058-6068.2005

Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Palm, C. J., et al. (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302, 842–846. doi: 10.1126/science.1088305

Yang, X. H., Tschaplinski, T. J., Hurst, G. B., Jawdy, S., Abraham, P. E., Lankford, P. K., et al. (2011). Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res.* 21, 634–641. doi: 10.1101/gr.109280.110

Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E., and Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2605–2610. doi: 10.1073/pnas.0509379103

Zhao, Q., Xiao, J., and Yu, J. (2012). An integrated analysis of lineage-specific small proteins across eight eukaryotes reveals functional and evolutionary significance. *Prog. Biochem. Biophys.* 39, 359–367. doi: 10.3724/SP.J.1206.2011.00290

Zuber, P. (2001). A peptide profile of the *Bacillus subtilis* genome. *Peptides* 22, 1555–1577. doi: 10.1016/S0196-9781(01)00492-2

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 11 October 2013; accepted: 27 November 2013; published online: 16 December 2013.*

*Citation: Su M, Ling Y, Yu J, Wu J and Xiao J (2013) Small proteins: untapped area of potential biological importance. Front. Genet.* ***4****:286. doi: 10.3389/fgene.2013.00286*

*This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.*